A method of synthesizing of an unvoiced speech signal

The present invention relates to the field of synthesizing of speech or music, and more particularly without limitation, to the field of text-to-speech synthesis.

The function of a text-to-speech (TTS) synthesis system is to synthesize speech from a generic text in a given language. Nowadays, TTS systems have been put into practical operation for many applications, such as access to databases through the telephone network or aid to handicapped people. One method to synthesize speech is by concatenating elements of a recorded set of subunits of speech such as demisyllables or polyphones. The majority of successful commercial systems employ the concatenation of polyphones. The polyphones comprise groups of two (diphones), three (triphones) or more phones and may be determined from nonsense words, by segmenting the desired grouping of phones at stable spectral regions. In a concatenation based synthesis, the conversation of the transition between two adjacent phones is crucial to assure the quality of the synthesized speech. With the choice of polyphones as the basic subunits, the transition between two adjacent phones is preserved in the recorded subunits, and the concatenation is carried out between similar phones.

Before the synthesis, however, the phones must have their duration and pitch modified in order to fulfil the prosodic constraints of the new words containing those phones. This processing is necessary to avoid the production of a monotonous sounding synthesized speech. In a TTS system, this function is performed by a prosodic module. To allow the duration and pitch modifications in the recorded subunits, many concatenation based TTS systems employ the time-domain pitch-synchronous overlap-add (TD-PSOLA) (E. Moulines and F. Charpentier, "Pitch synchronous waveform processing techniques for text-to-speech synthesis using diphones," Speech Commun., vol. 9, pp. 453-467, 1990) model of synthesis.

In the TD-PSOLA model, the speech signal is first submitted to a pitch marking algorithm. This algorithm assigns marks at the peaks of the signal in the voiced segments and assigns marks 10 ms apart in the unvoiced segments. The synthesis is made by a superposition of Hanning windowed segments centered at the pitch marks and extending from the previous pitch mark to the next one. The duration modification is provided by deleting or replicating some of the windowed segments. The pitch period modification, on

the other hand, is provided by increasing or decreasing the superposition between windowed segments.

Despite the success achieved in many commercial TTS systems, the synthetic speech produced by using the TD-PSOLA model of synthesis can present some drawbacks, mainly under large prosodic variations.

EP-0363233, US-A- 5,479,564, EP-0706170 disclose PSOLA methods. A specific example is also the MBR-PSOLA method as published by T. Dutoit and H. Leich, in Speech Communication, Elsevier Publisher, November 1993, vol. 13, N.degree. 3-4, 1993. The method described in document U.S. Pat. No. 5,479,564 suggests a means of modifying the frequency by overlap-adding short-term signals extracted from this signal. The length of the weighting windows used to obtain the short-term signals is approximately equal to two times the period of the audio signal and their position within the period can be set to any value (provided the time shift between successive windows is equal to the period of the audio signal). Document U.S. Pat. No. 5,479,564 also describes a means of interpolating waveforms between segments to concatenate, so as to smooth out discontinuities. When a noisy signal is to be synthesized by means of a known PSOLA method, the signal is repeated periodically. This way an unintended periodicity is introduced into the frequency spectrum. This is perceived as a metallic sound. This problem occurs for all noisy signals which do not have a fundamental frequency, such as unvoiced speech parts or music. An unvoiced speech part, like the "s" sound, has no pitch. The vocal chords are not moving as they do for a voiced sound. Instead, a noisy hiss-sound is produced by pushing air through a small opening between the vocal chords. Whisper is an example of speech containing only unvoiced parts. Where there is no pitch, there is no need to change it. However, it can be desirable to change the duration of an unvoiced speech part.

The present invention therefore aims to provide a method of synthesizing a signal which enables to modify the duration of unvoiced speech parts or music without introducing an unintended periodicity in the signal.

The present invention provides for a method of synthesizing a signal, in particular a noisy signal, based on an original signal. Further the present invention provides

for a computer program product for performing such a synthesis, as well as for a corresponding computer system, in particular, a text-to-speech system.

In accordance with the invention the required pitch bell locations of the signal to be synthesized are determined. This is done based on, for example, an assumed frequency of for example 100 Hz. This chosen frequency corresponds to a pitch period. The required pitch bell locations of the signal to synthesized are spaced apart on the time axis by intervals having the length of the pitch period. The required pitch bell locations are mapped onto the original signal to provide pitch bell locations in the domain of the original signal. The pitch bell locations in the domain of the original signal are randomly shifted. Preferably the randomization is performed by shifting the pitch bell locations in the original signal domain within +/- the pitch period.

In accordance with an embodiment of the invention the windowing is performed by means of a sine-window. The advantage of a sine-window is that it helps to reduce any residual periodicity. In particular using a sine-window is advantageous in that it ensures that the signal envelope in the power domain remains constant. Unlike a periodic signal, when two noise samples are added, the total sum can be smaller than the absolute value of any one of the two samples. This is because the signals are (mostly) not in-phase. The sine-window adjusts for this effect and removes the envelope-modulation.

In the following, preferred embodiments of the invention are described in greater detail by making reference to the drawings in which:

Fig. 1 is illustrative of a flow chart of an embodiment of the present invention,

Fig. 2 is illustrative of an example for synthesizing an unvoiced speech signal,

Fig. 3 is a block diagram of a preferred embodiment of a computer system.

The flow chart of Fig. 1 is illustrative an embodiment of the method of synthesizing a signal. In step 100 an original signal having a duration of y is provided. For example, the original signal is a natural speech signal containing unvoiced speech or a music signal having a noisy signal characteristic. Further a choice for a fundamental frequency f is made even though the original signal does not have such a fundamental frequency because of its noisy characteristics. The choice of a frequency f corresponds to a choice of a pitch period p. A convenient choice for a frequency f is between 50 Hz and 200 Hz, preferably 100 Hz. In

addition the desired duration x of the signal to be synthesized is inputted in step 100. In step 102 the pitch bell locations in the domain of the signal to be synthesized are determined in accordance with the choice of frequency f and pitch period p. This is done by dividing the time axis in the domain of the signal to synthesized into intervals of length p. In step 104 the pitch bell locations are mapped from the domain of the signal to be synthesized onto the domain of the original signal. When the duration x is longer than the duration y of the original signal this means that the pitch bell locations i in the domain of the original signal are spaced apart by intervals which are shorter than the pitch period p. In the opposite case the intervals between the pitch bell locations i in the domain of the original signal will be longer than the intervals between the pitch bell locations and the domain of the signal to be synthesized. In step 106 the pitch bell locations i in the domain of the original signal are randomized. This can be done by randomly shifting each of the pitch bell location i within an interval of +/- p around the original pitch bell location i. A pseudo random number generator can be utilized to perform this randomization. In step 108 the windowing is performed in the domain of the original signal. Preferably this is done by means of a sine-window which is applied on the randomized pitch bell locations i'; this way periodicity is further reduced. In step 110 the resulting pitch bells are overlapped and added in the domain of the signal to be synthesized which provides the synthesized signal.

Fig. 2 illustrates this signal synthesis by way of example. Time axis 200 is in the domain of the signal to be synthesized. The required duration x of the signal to be synthesized is one second in the example considered here. The assumed frequency f is 100 Hz, which corresponds to a pitch period p of 10 milliseconds. This means that the required pitch bell locations in the domain of the signal to be synthesized on time axis 200 are spaced apart by intervals of p = 10 milliseconds, i.e. the first pitch bell location is located at zero seconds on time axis 200, the next pitch bell location is at 10 milliseconds, the following at 20 milliseconds and so on. In other words the pitch bell locations in the domain of the signal to be synthesized are determined by points on the time axis 200 which are spaced apart by intervals of p starting at time zero. The pitch bell locations on time axis 200 are mapped onto time axis 202 in the domain of the original signal. The original signal has a duration of y = 0.5 seconds. As the duration y is smaller than the duration x of the signal to be synthesized this means that the pitch bell locations need to be "compressed" on time axis 202. As the duration y is half the duration x the intervals of the mapped pitch bell locations on the time axis 202 are spaced apart by p/2 instead of p. This means that the first pitch bell location i = 1 is at zero milliseconds on the time axis 202; the following pitch bell location i = 2 is at 5

milliseconds, the next pitch bell location i = 3 is at 10 milliseconds and so on. In other words the first pitch bell location at time zero milliseconds on the time axis 200 is mapped onto the pitch bell location i = 1 on the time axis 202 at zero milliseconds; the required pitch bell location at 10 milliseconds on the time axis 200 is mapped on the pitch bell location i = 2 at 5 milliseconds on the time axis 202; the required pitch bell location at 20 milliseconds on the time axis 200 is mapped onto the pitch bell location i = 3 at time 10 milliseconds on the time axis 202 and so on. Next the pitch bell locations i are randomized. This is illustrated in figure 2 with respect to the first pitch bell location i = 1 on the time axis 202. An interval of +/- p around zero milliseconds is defined on the time axis 202. Within this interval the pitch bell location i = 1 is randomly shifted. For the pitch bell location i = 1 the interval is between –10 milliseconds to +10 milliseconds on the time axis 202. In the example considered here this results in a randomized pitch bell location i' at 7.5 milliseconds on the time axis 202. At this position the original signal is windowed by means of a window function 204. Preferably the following window is used to provide a window function 204.

$$w[n] = \sin\left(\frac{\pi.(n+0.5)}{m}\right), \quad 0 \le n \ge m$$

Preferably the randomization of the pitch bell locations i is performed in accordance with the following formula:

$$i' = i + (R \times p)$$

Where i denotes the original pitch bell location on the time axis 202, i' is the new pitch bell location after the randomization, R is a random number between –1 and 1 and p is the pitch period. The result of the windowing of the original signal is a pitch bell. This pitch bell is placed at the first required pitch bell location within the domain of the signal to be synthesized on time axis 200 as illustrated in figure 2. This process is repeated with respect to all required pitch bells on the time axis. These pitch bells are added which yields the desired synthesized signal of length x.

Fig. 3 is illustrative of a block diagram of a computer system, such as a text-to-speech system. The computer system 300 has a module 302 for storing an original signal having a duration of y. Further the computer system 300 has a module 304 for storing a pre-selected frequency f or pitch p. Module 306 serves to determine required pitch bell locations of the signal to be synthesized based on the required duration x of the signal to be synthesized and the pre-selected frequency f or pitch p. Module 308 serves to map the required pitch bell locations in the domain of the signal to be synthesized onto the domain of

the original signal. This way the pitch bell locations i are determined as illustrated in the example of Fig. 2. Module 310 serves to randomize the pitch bell locations i. Module 310 is coupled to module 312 which provides random numbers for the randomization process. Module 314 serves to perform the windowing of the original signal on the randomized pitch bell locations i'. The resulting pitch bells are then overlapped and added in the domain of the signal to be synthesized by mean of module 316. This results in the synthesized signal of the desired duration y.

LIST OF REFERENCE NUMERALS:

time axis	200
time axis	202
window function	204
computer system	300
module	302
module	304
module	306
module	308
module	310
module	312
module	314
module	316